Supplement Material *for* Tuning Pre-trained Model via Moment Probing

Mingze Gao[1,2,†] Qilong Wang[1,*] Zhenyi Lin[1] Pengfei Zhu[1] Qinghua Hu[1] Jingbo Zhou[2]
[1]Tianjin Key Lab of Machine Learning, College of Intelligence and Computing, Tianjin University, China
[2]Business Intelligence Lab, Baidu Research, China
{gaomingze, qlwang, linzhenyi, zhupengfei, huqinghua}@tju.edu.cn, zhoujingbo@baidu.com

In the supplementary materials, we first give more descriptions about evaluation datasets. Then, we conduct experiments to further analyze our proposed Moment Probing (MP) method. Finally, the hyper-parameter details of our MP for tuning pre-trained models are provided.

1. Descriptions for Evaluation Datasets

We evaluate our methods on ten benchmarks, whose detailed descriptions are listed in Table S1.

FGVC. Following SSF [8], we employ five Fine-Grained Visual Classification (FGVC) datasets to evaluate the effectiveness of our methods, including CUB-200-2011 [15], NABirds [13], Oxford Flowers [11], Stanford Dogs [6], and Stanford Cars [2].

General Image Classification Datasets. We also validate the effectiveness of MP and MP_{+} on general image classification tasks. We use CIFAR-100, and ImageNet-1K as evaluation datasets, where CIFAR-100 contains 60,000 images with 100 categories and ImageNet-1K contains 1.28M training images and 50K validation images with 1,000 categories.

Out-of-Distribution Datasets. To verify the robustness of our MP, we conduct experiments on three out-of-distribution (OOD) datasets, including ImageNet-A (IN-A) [5], ImageNet-R (IN-R) [3] and ImageNet-C (IN-C) [4].

ImageNet-A 200 classes from 1,000 classes of ImageNet-1K and the real-world adversarial samples that make the ResNet model mis-classified are collected.

ImageNet-R contains rendition of 200 ImageNet-1K classes and 30,000 images in total.

ImageNet-C consists of the corrupted images, including noise, blur, weather, etc. The performance of model on ImageNet-C shows the robustness of model.

Dataset	#Classes	Train size	Val size	Test size
Fine-Grained Visual Classification (FGVC)				
CUB-200-2011	200	5,394	600	5,794
NABirds	55	21,536	2,393	24,633
Oxford Flowers	102	1,020	1,020	6,149
Stanford Dogs	120	10,800	1,200	8,580
Stanford Cars	196	7,329	815	8,041
General Image Classification Datasets				
CIFAR-100	100	50,000	-	10,000
ImageNet-1K	1000	1,281,167	50,000	150,000
Out-of-Distribution Datasets				
ImageNet-A	200	7,500		
ImageNet-R	200	30,000		
ImageNet-C	1000	$75 \times 50{,}000$		

Table S1: Details of evaluation datasets.

2. Evaluation on Harder Benchmark

In this section, to further assess effect of our MP, we conduct experiments on a more challenging (long-tailed and fine-grained) iNat2017 [14]. As shown in Table S2, proposed MP outperforms LP by 5.88% and 6.11% in top1 and top5 accuracy, respectively, while MP_{+} improves both SSF and Full Fine-tuning by a non-trivial gain (2.45% and 1.04%) in top-1 accuracy, verifying the effectiveness of both MP and MP_{+} on hard object recognition tasks.

3. Visualization of Attention Maps

Here we analyze our methods by visualizing learned attention maps, while comparing with linear probing (LP) and full fine-tuning methods. Figure S1 gives some examples sampled ImageNet-1K, where we have the following obser-

[†] This work was done when Mingze Gao was an intern at Baidu Research. * Corresponding author

Method	Top-1 Acc. (%)	Top-5 Acc. (%)
Linear Probing	56.61	80.01
MP (Ours)	$62.49_{(5.88)}$	$86.12_{(6.11)}$
SSF [8]	66.30	88.29
Full Fine-tuning	67.71	88.81
MP_{+} (Ours)	**68.75**	**89.04**

Table S2: Comparison of different tuning methods on iNat2017 dataset, where ViT-B/16 pre-trained on ImageNet-21K is used as basic backbone.

vations: (1) LP generally fails to focus on appropriate regions, while our MP and MP_{+} methods can accurately capture key information, as shown in the third and fifth rows of figure S1. (2) Full fine-tuning method may result in a deterioration of generalization ability during the fine-tuning process, and is unable to capture appropriate regions. In contrast, as shown in the first and third rows of Figure S1, both MP and MP_{+} can correctly attend to important regions and have strong robustness.

4. Training Details

In this work, we apply the proposed MP to ViT-B/16 [1], Swin-B [9], ConvNeXt-B [10] and AS-MLP-B [7] under fine-tuning and few-shot settings. Here we show the details of optimization policy and hyper-parameter settings in Table S3. For fine-tuning ViT-B/16, Swin-B/16, ConvNeXt-B on ImageNet-1K and AS-MLP-B on CIFAR-100, we follow the configurations in [8, 1]. For few-shot learning on ImageNet-1K, we refer to the configurations in [16, 12] where ViT-B/16 pretrained on ImageNet-21K as backbone.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[2] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 1

[3] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[4] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 1

[5] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

[6] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization*, 2011. 1

[7] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. As-mlp: An axial shifted mlp architecture for vision. In *International Conference on Learning Representations*, 2022. 2

[8] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems*, 2022. 1, 2

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 2

[10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[11] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 1

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 2

[13] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 1

[14] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1

[15] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[16] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *European Conference on Computer Vision*, 2022. 2

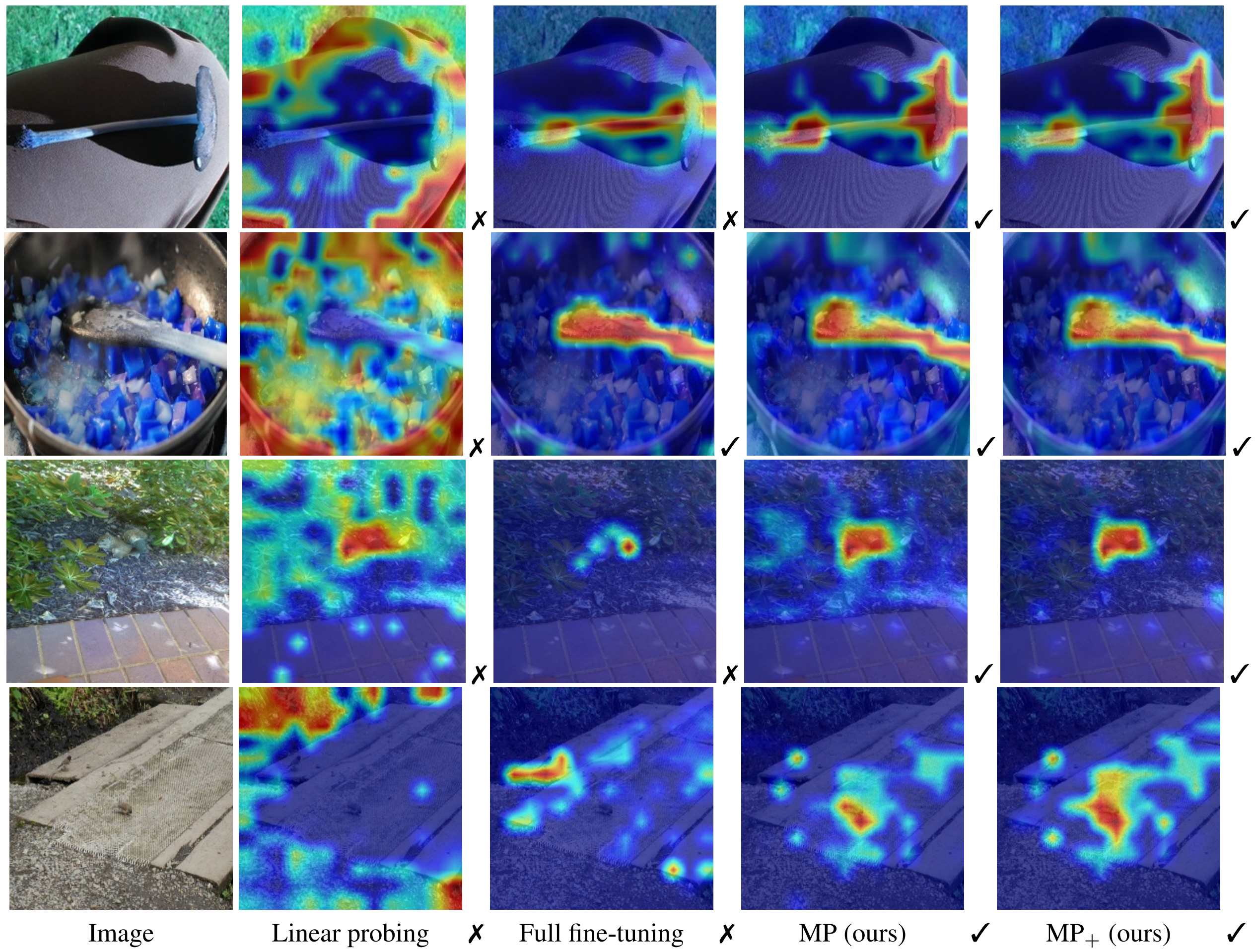

Figure S1: Visualization of attention maps. From left to right, each column shows the original images and attention maps achieve by linear probing, full fine-tuning, our MP and MP_+. All models are pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K using the ViT-B/16 model.

	Fine-tuning settings								Few-shot settings	
Model	ViT-B/16		Swin-B		ConvNeXt-B		AS-MLP-B		ViT-B/16	
Methods	Full	LP / MP	Full	LP / MP	Full	LP / MP	Full	LP / MP	Full	LP / MP
Batch size	256	256	256	256	256	256	256	256	32	32
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Scheduler	cosine	cosine	cosine	cosine	cosine	cosine	cosine	cosine	cosine	cosine
Momentum	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Epochs	30	30	30	30	30	30	100	100	20	20
Base learning rate	1e-4	1e-4	5e-5	5e-4	5e-5	1e-3	5e-5	1e-3	3e-4	2e-5
Min learning rate	1e-8	1e-8	5e-8	5e-8	5e-8	5e-8	5e-8	1e-8	1e-8	1e-8
Warmup epochs	5	5	5	5	5	5	10	10	0	0
Warmup learning rate	1e-7	1e-7	5e-7	5e-7	5e-7	5e-7	5e-7	1e-7	-	-
Weight decay	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Drop path	0.2	0.1	0.2	0.1	0.2	0.1	0	0	0	0

Table S3: Details of hyper-parameter settings of our MP methods, which involve fine-tuning and few-shot settings in various deep architectures on ImageNet-1K (CIFAR-100 for AS-MLP), where linear probing, MP, and MP_+ methods all use the same settings as shown in column (LP / MP), while full fine-tuning settings are shown in column (Full).